Genetics in Medicine Open (2024) 2, 101830

ELSEVIER

Genetics in Medicine OPEN
An Official Journal of the ACMG

www.journals.elsevier.com/genetics-in-medicine-open

ARTICLE

The impact of the Turkish population variome on the genomic architecture of rare disease traits

Zeynep Coban-Akdemir[1,2], Xiaofei Song[1,3], Francisco C. Ceballos[4], Davut Pehlivan[1,5], Ender Karaca[1,6,7], Yavuz Bayram[1,8,9], Tadahiro Mitani[1], Tomasz Gambin[10], Tugce Bozkurt-Yozgatli[2,11], Shalini N. Jhangiani[12], Donna M. Muzny[12], Richard A. Lewis[1,13,14], Pengfei Liu[1], Eric Boerwinkle[2,12], Ada Hamosh[15], Richard A. Gibbs[1,12], V. Reid Sutton[1,16], Nara Sobreira[15], Claudia M.B. Carvalho[1,17], Chad A. Shaw[1,18], Jennifer E. Posey[1], David Valle[15], James R. Lupski[1,12,13,16,*]

ARTICLE INFO

Article history:
Received 28 July 2023
Received in revised form
3 February 2024
Accepted 7 February 2024
Available online 14 February 2024

Keywords:
Admixture
Consanguinity
Genomic architecture of rare disease traits
Runs of homozygosity
Turkish population

ABSTRACT

Purpose: The variome of the Turkish (TK) population, a population with a considerable history of admixture and consanguinity, has not been deeply investigated for insights on the genomic architecture of disease.

Methods: We generated and analyzed a database of variants derived from exome sequencing data of 773 TK unrelated, clinically affected individuals with various suspected Mendelian disease traits and 643 unaffected relatives.

Results: Using uniform manifold approximation and projection, we showed that the TK genomes are more similar to those of Europeans and consist of 2 main subpopulations: clusters 1 and 2 (N = 235 and 1181, respectively), which differ in admixture proportion and variome (https://turkishvariomedb.shinyapps.io/tvdb/). Furthermore, the higher inbreeding coefficient values observed in the TK affected compared with unaffected individuals correlated with a larger median span of long-sized (>2.64 Mb) runs of homozygosity (ROH) regions (P value = 2.09e-18). We show that long-sized ROHs are more likely to be formed on recently configured haplotypes enriched for rare homozygous deleterious variants in the TK affected compared with TK unaffected individuals (P value = 3.35e-11). Analysis of genotype-phenotype correlations reveals that genes with rare homozygous deleterious variants in long-sized ROHs provide the most comprehensive set of molecular diagnoses for the observed disease traits with a systematic quantitative analysis of Human Phenotype Ontology terms.

Conclusion: Our findings support the notion that novel rare variants on newly configured haplotypes arising within the recent past generations of a family or clan contribute significantly to recessive disease traits in the TK population.

The Article Publishing Charge (APC) for this article was paid by Zeynep Coban-Akdemir.

Zeynep Coban-Akdemir, Xiaofei Song, and Francisco C. Ceballos contributed equally.

*Correspondence and requests for materials should be addressed to James R. Lupski, Baylor College of Medicine, Molecular and Human Genetics, One Baylor Plaza, Room 604B, MS: BCM225, Houston, TX 77030-3411. *Email address:* jlupski@bcm.edu

Affiliations are at the end of the document.

doi: https://doi.org/10.1016/j.gimo.2024.101830

Introduction

Population genetics has been applied for many decades as a means to uncover loci that contribute to human disease by the genome-wide analyses of single-nucleotide variation (SNV, formerly SNP). These SNVs are common variants, frequently shared among geographically distinct populations, and enable a fine-scale mapping of recombination to yield potential disease-contributing loci. This genome-wide association study approach has been applied largely to the study of common, complex conditions for which many have hypothesized that a combination of common variants each with small effect size can influence genetic susceptibility to disease expression.[1,2]

More recent studies, however, have uncovered a rare variant contribution to some apparently complex conditions (ie, chronic kidney disease, scoliosis, arthrogryposis, developmental delay, and intellectual disability), suggesting that a substantial proportion of seemingly common, complex conditions may represent, in fact, a combination of individually rare, Mendelian disease traits.[3-7]

The role of ultra-rare variants in recessively inherited conditions has been less studied, although novel variants contributing to recessive disease have been reported in non-European origin population-specific cohorts.[8-10] To investigate comprehensively the role of new variants in rare recessive disease traits, we created a population-specific rare disease cohort and rare variant database against which to measure population-based frequencies in the context of population genetic substructure.

The transmission of traits, genes, and variant alleles from one generation to the next may result in identity-by-descent (IBD) at a locus in a population characterized by consanguinity or a founder effect due to a historical population bottleneck or geographic isolation. Experimentally, evidence for IBD in an individual genome is suggested by the presence of runs of homozygosity (ROH) not accompanied by copy-number variation; ie, copy-number-neutral gene dosage at a locus. Analysis of ROH regions in an individual genome can be used to prioritize potential pathogenic variations at a gene locus and may unveil possible genetic susceptibility to underlying disease.[11-13]

The inter-individual variation in ROH number and total length has also been shown to contribute to the genetic architecture of complex traits and rare diseases.[14-17] Therefore, analysis of ROH regions can be an adjuvant analytical tool to address the genetic architecture of disease. For instance, homozygosity mapping has been a robust genetic approach for the identification of biallelic variants causing autosomal recessive disease traits. For the identified "disease gene," the locus resides in ROH blocks shared among affected individuals.[18-23] This approach was used successfully to map disease genes in consanguineous families with heterogeneous neurological disorders,[24] multisystem disorders,[25,26] and inborn errors of metabolism.[8-10] Furthermore, analysis of rare homozygous deleterious variants in ROH regions can elucidate the genetic heterogeneity and the molecular mechanisms underlying Mendelian traits due to either the contribution of multi-locus variation or oligogenic/polygenic recessive effects[3,7,27-33] and even provide a molecular genetic explanation for trait penetrance, variability of expression of disease,[34] and evidence for potential modifying gene loci.[35-37]

Both the frequency and the degree of relatedness for consanguineous marriages within populations vary in different geographic regions and countries around the globe; the highest documented frequency, 60% to 76%[38] of unions, is recorded in Pakistan. Although consanguineous families facilitate genetic locus mapping and disease gene discovery, many populations with a relatively high coefficient of consanguinity, such as the Turkish (TK) population,[39,40] have not been investigated deeply. The population genetics of Turkey present peculiar features compared with other Middle East countries. Besides having a high reported rate of consanguinity, 20.1%,[41] Turkey also is described often as both a geographic and a social "bridge" between Asia and Europe, an important hub of both ancient and contemporary population migration. Population substructure studies in the TK population potentially can provide insights about the effects of high admixture and a relatively increased rate of consanguinity to impact the genomic architecture of disease.

To investigate the influences of both admixture and consanguinity on population genetic variation and the genetic architecture of disease, we studied a sizeable TK cohort (1416 personal genomes) with a considerable amount of consanguinity and admixture. We performed exome sequencing (ES) and family-based genomic analysis on 1416 TK individuals consisting of 773 unrelated clinically affected individuals with a wide variety of suspected Mendelian disease traits and 643 unaffected relatives. We performed an unbiased exome variant analysis that would enable a rare variant family-based genomics approach to elucidate the molecular etiology and define gene loci potentially contributing to their clinically observed disease traits. We further carried out systematic genomic and phenotypic analysis of this population, by genomic variant detection and with a structured ontology of human phenotype ontology (HPO) terms, which might reveal key features of the TK population variome and substructure that affect the genetic architecture of disease in this TK cohort.

Material and Methods

Experimental model and participant details

We recruited 1416 unrelated TK individuals (669 females and 747 males) in the Baylor Hopkins Center for Mendelian Genomics (BHCMG) cohort (data freeze: December 2011-October 2020) after all relevant subjects or legally authorized representatives provided written informed consent for the use of their DNA and personal genomes for the

identification of potential disease-contributing variants and for broad data sharing. Peripheral blood was collected from affected individuals, parents, and unaffected relatives if available. Genomic DNA was extracted from blood leukocytes according to standard procedures. All genomic studies were performed on DNA samples isolated from blood.

Materials availability

This study did not generate unique biological or chemical reagents.

ES and annotation

ES was performed at the Human Genome Sequencing Center (HGSC) at Baylor College of Medicine (BCM) through the BHCMG initiative. With 1 μg of DNA, an Illumina paired-end pre-capture library was constructed according to the manufacturer's protocol (Illumina Multiplexing_SamplePrep_Guide_1005361_D) with modifications described in the Baylor College of Medicine-HGSC Illumina Barcoded Paired-End Capture Library Preparation protocol. Pre-capture libraries were captured into 4-plex library pools and hybridized in solution to the HGSC-designed Core capture reagent (52 Mb, NimbleGen) or 6-plex library pools with the custom VCRome 2.1 capture reagent (42 Mb, NimbleGen), according to the manufacturer's protocol (NimbleGen SeqCap EZ Exome Library SR User's Guide) with minor revisions. The sequencing was performed in paired-end mode with the Illumina HiSeq 2000 platform or Illumina NovaSeq 6000 platforms. Data were aligned to GRCh37/hg19 with BWA-aln (for data generated on HiSeq 2000) or BWA-mem (for data generated on NovaSeq). Sequence analysis was performed with the HGSC Mercury analysis pipeline (https://www.hgsc.bcm.edu/software/mercury),[42,43] which moves data through various analysis tools from the initial sequence generation to annotated variant calls (SNVs and intra-read insertion/deletions; ie, indels). Variants were called with ATLAS2 or xATLAS[44] and the Sequence Alignment/Map (SAMtools) suites and annotated with an in-house-developed Cassandra[45] annotation pipeline that uses Annotation of Genetic Variants (ANNOVAR)[46] and additional tools and databases, including ExAC (http://exac.broadinstitute.org), gnomAD (https://gnomad.broadinstitute.org), the Greater Middle Eastern (GME) variome (https://annovar.openbioinformatics.org/en/latest/user-guide/filter/), and the Atherosclerosis Risk in Communities database (http://drupal.cscc.unc.edu/aric/). In our comparative analysis to the other data sets, GME and gnomAD, to ensure harmonization across all the diverse capture designs of the data sets, we only used the variants that were exonic or splicing.

Phenotypic characterization of the BHCMG cohort

We used the PhenoDB database to collect and store the information of clinical features, pedigree structures, and self-reported consanguinity levels. Computational analyses of phenotyping data were performed by HPO terms analyses described as a phenotypic similarity score with the R package ontology Similarity.[29,47,48]

Obtaining a final set of unrelated individuals in the TK cohort

To minimize overrepresentation of individuals from the same family and to identify mutually unrelated individuals in the TK cohort, we used the PC-AiR function in the R GENESIS package, which measures pairwise kinship coefficients and ancestry divergence to identify an ancestry representative subset of mutually unrelated individuals. We removed the individuals with a close relationship from the analysis by a threshold of kinship coefficient as 0.044 (third-degree relationship or more) from both the TK affected and unaffected cohorts. Removing individuals with a close relationship from the analysis resulted in a final set of 1416 unrelated TK individuals: 773 unrelated affected and 643 unrelated unaffected participants.

Population substructure analysis

We investigated the population structure of this TK cohort along with the African, East Asian, South Asian, and European population samples from the 1000 Genomes Project phase 3 release data, including 2504 individuals' genotypes in total (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). First, genotypes for 31 individuals who have a biological relationship with the 2504 samples were removed from the analysis. Although 2 capture designs were used (HGSC-designed Core capture reagent [52 Mb, NimbleGen] or 6-plex library pools with the custom VCRome 2.1 capture reagent [42 Mb, NimbleGen]) throughout the BHCMG CMG initiative, the variants overlapping with both of the designs were used for all of the analyses performed including the admixture analyses.

Variants falling out of the HGSC-designed Core capture design and VCRome 2.1 capture design in the 1000 Genomes Project data were also filtered out from the analyses. A pruned subset of the remaining polymorphic single-nucleotide variants (SNVs) that are in approximate linkage equilibrium of each other (N = 89,379 for the TK affected participants and N = 86,049 for the TK unaffected participants) was used for the Uniform manifold approximation and projection (UMAP). UMAP was performed with the umap R package. For the admixture analysis, we ran the unsupervised ADMIXTURE algorithm by k = 5 clusters that outputted the minimal cross validation error.

Estimation of inbreeding coefficient values

The coefficient of inbreeding of an individual represents the probability that 2 alleles at any randomly chosen locus in an individual are identical by descent. The inbreeding coefficient values of the TK cohort were estimated from the ES

data with plink –het function filtering the variants with MAF ≥ 0.05.

Identifying and analyzing ROH segments from ES data

We detected ROH regions from unphased ES data as Absence of Heterozygosity genomic intervals with BafCalculator (https://github.com/BCM-Lupskilab/BafCalculator).[32] To call ROH regions with BAFCalculator, we extracted all the high-quality SNVs residing in the capture region (mostly exonic regions) available from the variant call format file of each single individual's exome. For those SNVs, we extracted a B-allele frequency (ie, variant reads/total reads ratio); then, we transformed this ratio by subtracting 0.5 and taking the absolute value for each data point. Transformed B-allele frequency data were processed with Circular Binary Segmentation implemented in the DNAcopy R Bioconductor package[49,50] to call the ROH regions. This algorithm merges the consecutive exon calls; therefore, in this way, we can detect ROH regions all over the genome that are of size ranging from hundreds of Kb to a few Mb, including a number of genes. To test the false-positive and false-negative rate of exome data and our algorithm, BafCalculator, we ran an independent analysis. We ran the BafCalculator to call ROH regions from an independent data set consisting of 929 samples with both genome sequencing (unphased GS) and high-resolution phased array data available in the Human Genome Diversity panel (https://www.internationalgenome.org/data-portal/data-collection/hgdp). Then, we compared ROH regions identified by the BafCalculator with the GS data with the ROH regions detected through high-resolution array in those 929 samples. The BafCalculator algorithm was further optimized by the segmentation mean (seg.mean) parameter (an absolute measure of average homozygosity rate of a putative ROH call). Cross-referencing the ROH regions identified by the fine-tuned BafCalculator to the array data ROH calls showed a positive predictive value of 90% and true-positive rate of 72% when the seg.mean parameter = 0.47 (Supplemental Figure 1). After this analysis, we also took the exome portions of the genome data in those samples and ran the BafCalculator with only those regions. Then, we compared FROH (the total size of ROHs ≥ 1.5 Mb) estimates identified from those regions (FROH_ES) to the FROH estimates identified from the genome sequencing data (FROH_GS). These analyses revealed that, when the seg.mean = 0.47, this provides a nearly perfect correlation (0.98) between FROH_ES and FROH_GS, and that ES performs similarly to GS (Supplemental Figure 2). In summary, segments with the mean signal > 0.47 and number of marks ≥ 10 were classified as ROH regions.

The calculated ROH intervals from BafCalculator could represent individual genomic/gene loci, resulting in ROH for diploid alleles that can occur by (1) IBD, (2) uniparental disomy (UPD),[51] or (3) a large deletion copy number variant (CNV). To exclude the ROH blocks that could be caused by genomic overlapping of a common variant deletion CNV, we first identified deletion CNVs through eXome-hidden Markov model (XHMM).[52] We further intersected ROH segments and potential deletion CNVs with BEDTools[53] and then retained only ROH regions overlapping less than 50% of their size with a variant deletion CNV. We grouped ROH regions into 3 size categories, applying Gaussian-mixture modeling from the MClust function in mclust R package into 3 length classes: long-sized genomic intervals or ROH blocks (>2.64 Mb), medium-sized ROH blocks (0.671-2.64 Mb), and short ROH blocks (0.210-0.671 Mb).

To control for the variable rates across different genomic regions and among different individuals, for each individual i and ROH region category $r, r \in \{total-ROH, non-ROH, long-sized\ ROH, medium-sized\ ROH, short-sized\ ROH\}$, we computed a variant density $f_{i,r}$

$$f_{i,r} = \frac{N^d_{i,r}}{N^s_{i,r}}$$

in which $N^d_{i,r}$ is the count of rare homozygous deleterious or likely damaging variant alleles (variants above a certain CADD score (≥15) that are located within a region r, whereas $N^s_{i,r}$ is the count of synonymous variants in a region r.

Generation of an objective score for phenotypic similarity comparison

We applied 2 R packages, OntologyIndex and ontologySimilarity, to measure the phenotypic similarity between 2 sets of HPO terms; each set of terms was associated with a patient's clinical features recorded in PhenoDB (https://phenodb.org/).[54] To assess the ability of a "candidate disease-contributing gene" to explain a patient's clinical features, we applied a previously published method, in which first a MICA (most informative common ancestor) matrix was calculated for each pair of HPO terms, and then a Resnik score was calculated for 2 sets of HPO terms.[29,47,48]

First, we calculated the information content (IC) for HPO term t by

$$IC(t) = -\log\left(f(t)\right)$$

in which $f(t)$ is the frequency of term t observed in all of the OMIM entries. The similarity of term i and term j is calculated with a Resnik method:

$$R_{ij} = IC\left(MICA(t_i, t_j)\right)$$

in which the Resnik score, R_{ij}, is determined by the IC of the

most informative common ancestor (MICA) of term i and term j. Next, we defined the phenotypic similarity score Sim for 2 HPO sets l_1 and l_2 as follows:

$$l_1 = \{t_{11}, t_{12}, \ldots, t_{1m}\}$$

$$l_2 = \{t_{21}, t_{22}, \ldots, t_{2n}\}$$

$$Sim(l_1, l_2) = \frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{m} \max_{1 \le j \le n} R_{ij}, \frac{1}{n}\sum_{j=1}^{n} \max_{1 \le i \le n} R_{ij}\right)$$

The main features of known human disease genes were summarized in OMIM (https://www.omim.org/) in the format of both plain texts and clinical synopses, and the associated HPO terms for each OMIM entry were annotated manually by the HPO-team (https://hpo.jax.org/app/data/annotationsWe adapted this method to measure the phenotypic similarity between a patient's phenotypes and a list of disease genes. For each list of the tested disease genes, we calculated a z-score performing 1000 simulations; in each simulation, we computed a similarity score between the patient's clinical features and the associated HPO terms of a randomly selected disease gene list, which has a same number of genes as the tested disease gene list. Further, to compare the contribution of disease gene lists across different genomic regions with the explanation of a patient's clinical phenotypic features, we computed a ratio of the similarity score calculated for a subset of disease genes, eg, genes located in long-sized ROH regions to the similarity score calculated for all the associated disease genes in a patient's genome.

Statistical analyses

We performed the statistical analyses with R version 3.3.3. We compared pairwise differences in the average values of estimated inbreeding coefficient values (F), homozygous rare deleterious variant burden (density), and total, median length, and count of ROHs in 2 participant groups, TK affected and TK unaffected participants by the Wilcoxon rank-sum one-tailed test. There were more than one 2-way comparisons in Figures 3B, 4A, 4B, 5B, and 6 and Supplemental Figures 1, 4, 7, 8, and 11, and the overall expected false-positive rate of the analysis was controlled by adopting a Bonferroni adjustment dividing by 0.05/(the number of 2-way comparisons). Bar plots, box plots, pie charts, and scatter plots were generated by ggplot2 data visualization R package, and stat_pvalue_manual function in the ggpubr R package added P values and significance levels to those plots. Ddply function in plyr CRAN R package reported the summary statistics of estimated F values, homozygous rare deleterious variant burden (density), and total, median length, and count of ROHs in 2 participant groups, TK affected and TK unaffected participants.

Results

The fine-scale population substructure of the TK cohort

To study finer-scale population substructure of the TK individuals in comparison with the African, East Asian, European, and South Asian population samples from the 1000 Genomes Project, we performed the UMAP dimension reduction method. The first and second main UMAP components separated the samples from African, East Asian, European, South Asian, and TK populations. These studies showed that the TK genomes were distinct from the African, East Asian, and South Asian populations, but closely clustered with the variome of European samples and consist of 2 main subpopulations, cluster 1 ($N = 235$) and cluster 2 ($N = 1181$) individuals, (Figure 1A and B).

To evaluate the population substructure of the TK individuals, we ran the unsupervised ADMIXTURE algorithm by $k = 5$ clusters to minimize cross validation error. The admixture analysis results revealed that, compared with the first cluster ($N = 235$), the second cluster ($N = 1181$) demonstrated a higher fraction of East Asian, European, South Asian (Wilcoxon test one-tailed P values = 1.39e-4, 3.47e-47, and 4.61e-36), and a lower fraction of other (Wilcoxon test one-tailed P values = 4.61e-36) ancestry but do not differ significantly from each other in the African ancestry component (P values = .172; Supplemental Figure 3). To note, there are a few TK affected and unaffected participants closely clustered with the African and East Asian population samples from the 1000 Genomes project. This likely reflects the increased number of people from all over the world that have recently migrated to Turkey during the last decade.

Comparison of the TK cohort variome with control for variant database bias

We next questioned how the population substructure was represented in the TK variome by examination of the characteristics of the distinct 3,101,517 variants present in TK individuals. We found that 675,693 and 1,893,512 variants are present in the TK unaffected cluster 1 or 2 individuals, and 1,202,217 and 1,961,838 are present uniquely in the TK affected cluster 1 or 2 individuals, respectively (Figure 2A). We then surveyed all the exonic variants of the TK variome in control variant databases, including the genome aggregation database (gnomAD, https://gnomad.broadinstitute.org)[55] version 4 (v4) data set and the GME variome.[56] Regarding the variome in the TK unaffected participants ($N = 643$), the

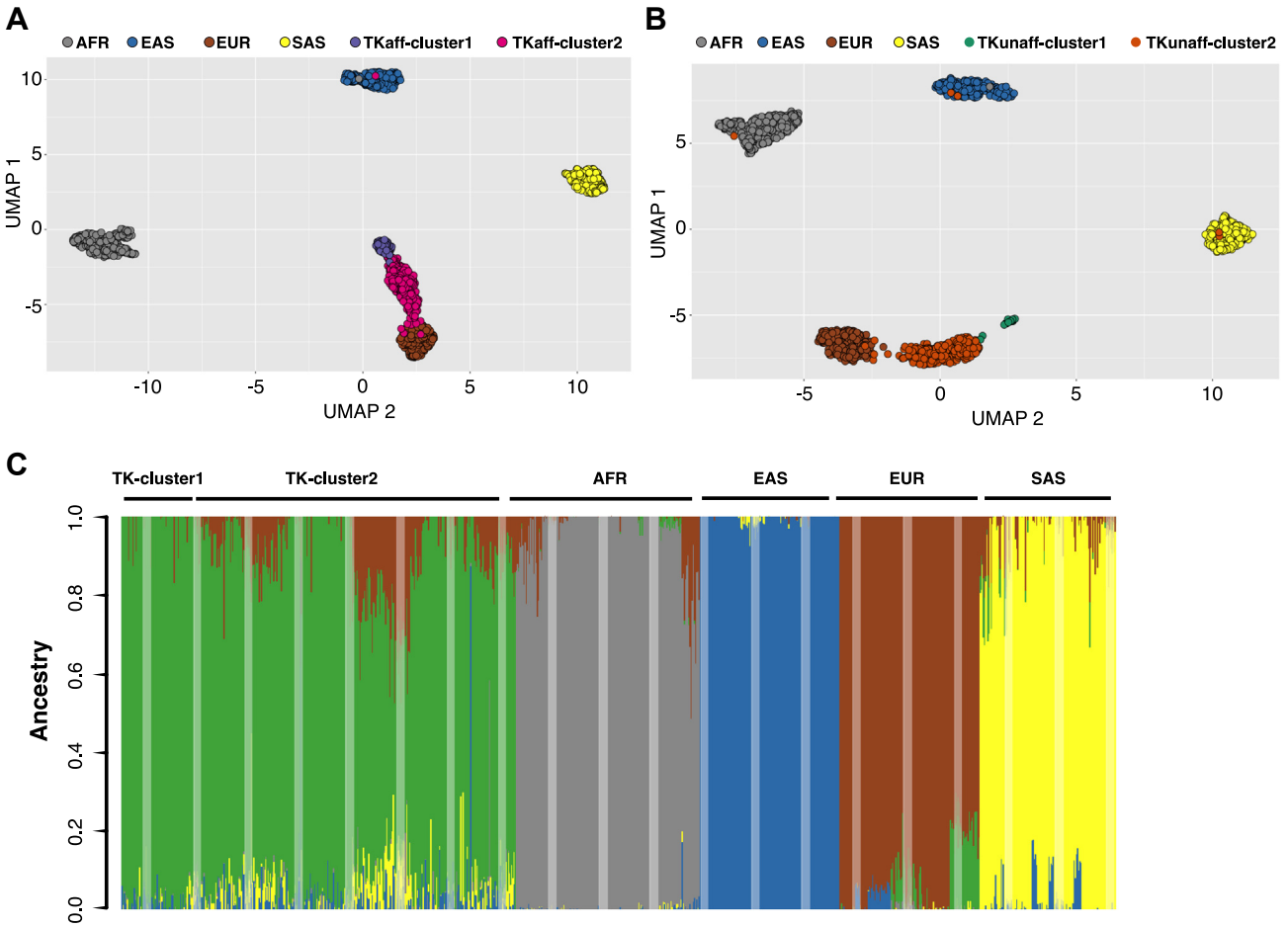

Figure 1 Fine-scale population substructure of TK cohort. Scatter plots display uniform manifold approximation and projection (UMAP) analysis that compares the population structure among the (A) TK affected participants ($N = 773$, colored in purple and dark pink) and (B) TK unaffected participants ($N = 643$, colored in dark green and dark orange) of the BHCMG cohort to the African (AFR) ($N = 661$, colored in gray), East Asian (EAS) ($N = 504$, dark blue), European (EUR) ($N = 503$, brown), and South Asian (SAS) ($N = 489$, yellow) population samples from the 1000 Genomes project. (C) The bar plot demonstrates the results of the admixture analysis performed through the unsupervised ADMIXTURE algorithm ($k = 5$ clusters) quantifying the fraction of ancestry proportions contributed by the African (gray), East Asian (dark blue), European (brown), South Asian (yellow), and Other (light brown) for the TK cluster 1 and cluster 2 individuals. To note, there are a few TK affected and unaffected participants closely clustered with the African and East Asian population samples from the 1000 Genomes project. This likely reflects the increased number of people from all over the world that have recently migrated to Turkey during the last decade.

overall comparison of all distinct SNVs identified showed that 55% and 54% of the unique variants in cluster 1 and cluster 2 individuals, respectively, were present in the gnomAD v4 variome. Intriguingly, only 10% and 5% of the cluster 1 and 2 variants, respectively, in the TK unaffected participants were present in the GME variome, underscoring the necessity of population-matched control databases to assess accurately variant minor allele frequency (Figure 2D and E, Supplemental Figure 4C and D).

We also performed similar analyses for the variome of TK affected participants ($N = 773$). These studies demonstrated that 49% and 53% of cluster 1 and cluster 2 variants, respectively, were represented in the gnomAD v4 database. Of note, only either 7% or 5% of the cluster 1 and cluster 2 variants were represented in the GME variome (Figure 2A and B, Supplemental Figure 4A and B). Aggregate TK exome variant data from clusters 1 and 2 are available publicly for analysis through a TK variome database (https://turkishvariomedb.shinyapps.io/tvdb/).

Observation of higher estimated inbreeding coefficient values in the TK participants

To obtain a more objective experimental measure of the consanguinity level, we estimated the inbreeding coefficient (F) values from ES data of the TK affected ($N = 773$) and unaffected participants ($N = 643$) compared with the African ($N = 661$), East Asian ($N = 504$), European ($N = 503$), and South Asian ($N = 489$) population samples from the 1000 Genomes Project. These analyses revealed that the estimated F values in the TK unaffected participants were significantly higher with mean values of 0.017 and 0.029 in

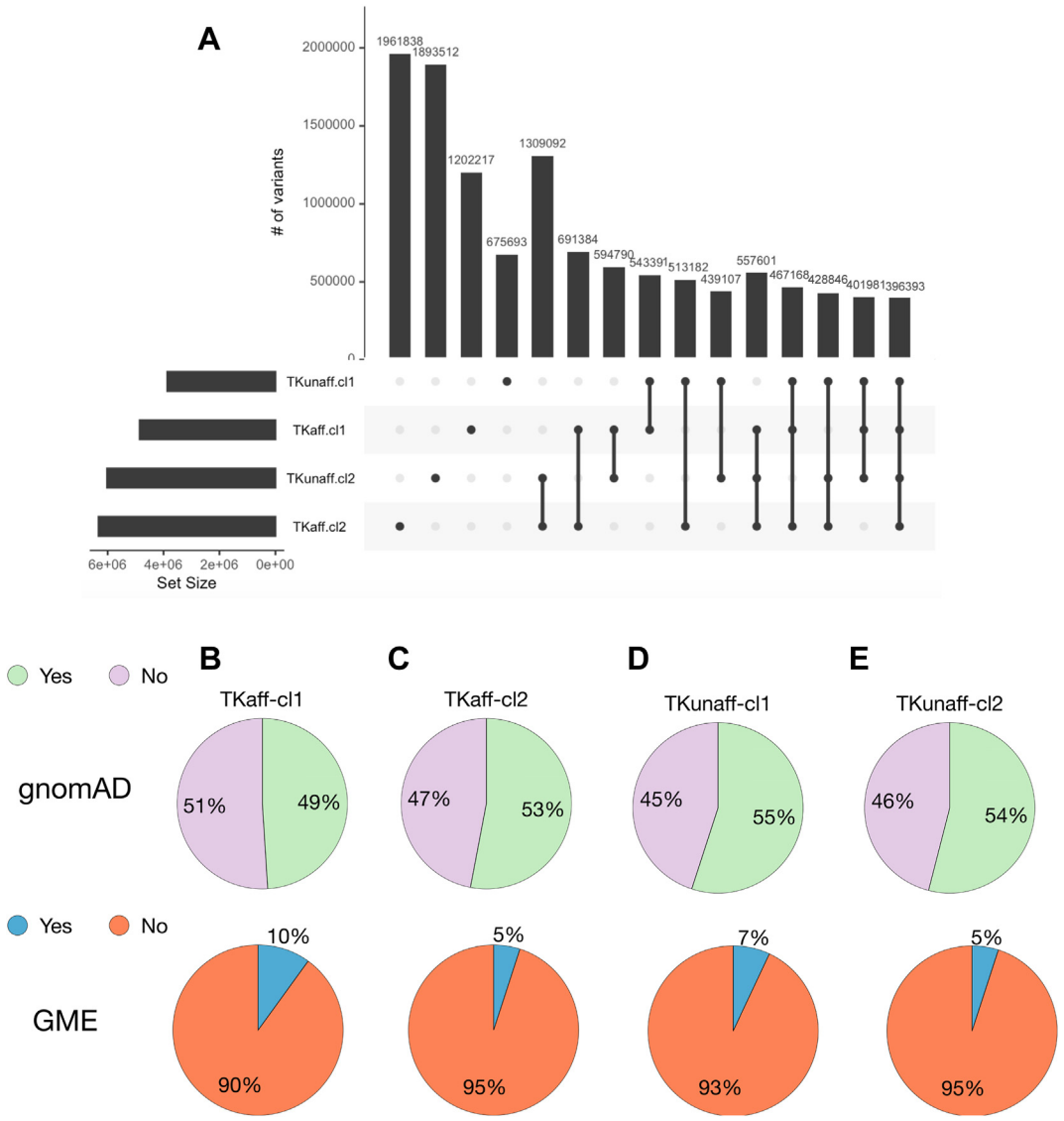

Figure 2 Comparison of the TK cohort variome with control databases. A. The Upset plot depicts the number of variants present in the TK affected cluster 1 (TKaff-cl1), affected cluster 2 (TKaff-cl2), unaffected cluster 1 (TKunaff-cl1), and unaffected cluster 2 (TKunaff-cl2) participants ($N = 3{,}101{,}517$ variants in total). B-E. The pie charts represent all distinct variants in the TK cohorts and whether they are present ("yes") or absent ("no") in gnomAD v4 data set (upper panel) and the Greater Middle Eastern Variome (lower panel) by affected status and cluster: (B) unaffected cluster 1, (C) unaffected cluster 2, (D) affected cluster 1, and (E) affected cluster 2.

cluster 1 and cluster 2 individuals, respectively, when compared with African (0.004), East Asian (−0.001), European (−0.0006), and South Asian (0.012) persons (Figure 3A). We showed further that the estimated F values of TK affected cluster 1 and cluster 2 individuals (mean = 0.055 vs 0.052) were increased significantly compared with the TK unaffected cluster 1 and cluster 2 individuals, respectively (Wilcoxon test one-tailed *P* values 2.3e-6 and 2.2e-16, Figure 3A and B). On the other hand, we did not observe any significant cluster-specific differences in the estimated F values of either TK affected (Wilcoxon test one-tailed *P* value .34) or TK unaffected individuals (Wilcoxon test one-tailed *P* value .11, Figure 3B). Therefore, we merged these 2 clusters for further analyses. Our analysis also showed that the measured genomic inbreeding coefficients—the fraction of the genome covered by ROHs > 1.5 Mb (FROH)—are 0.048 and 0.030 on average in the TK affected and unaffected participants, respectively, and these are nearly 1-1 proportional to the average estimated F values of the TK affected (0.053) and unaffected participants (0.028). These findings support the contention that the excess of homozygosity in the TK genomes was shaped mostly by ROHs (Figure 3C).

Enrichment of long-sized ROH genomic regions in TK cohort individuals

Then we hypothesized that higher F values measured in the TK cohort correlate with an increased length and

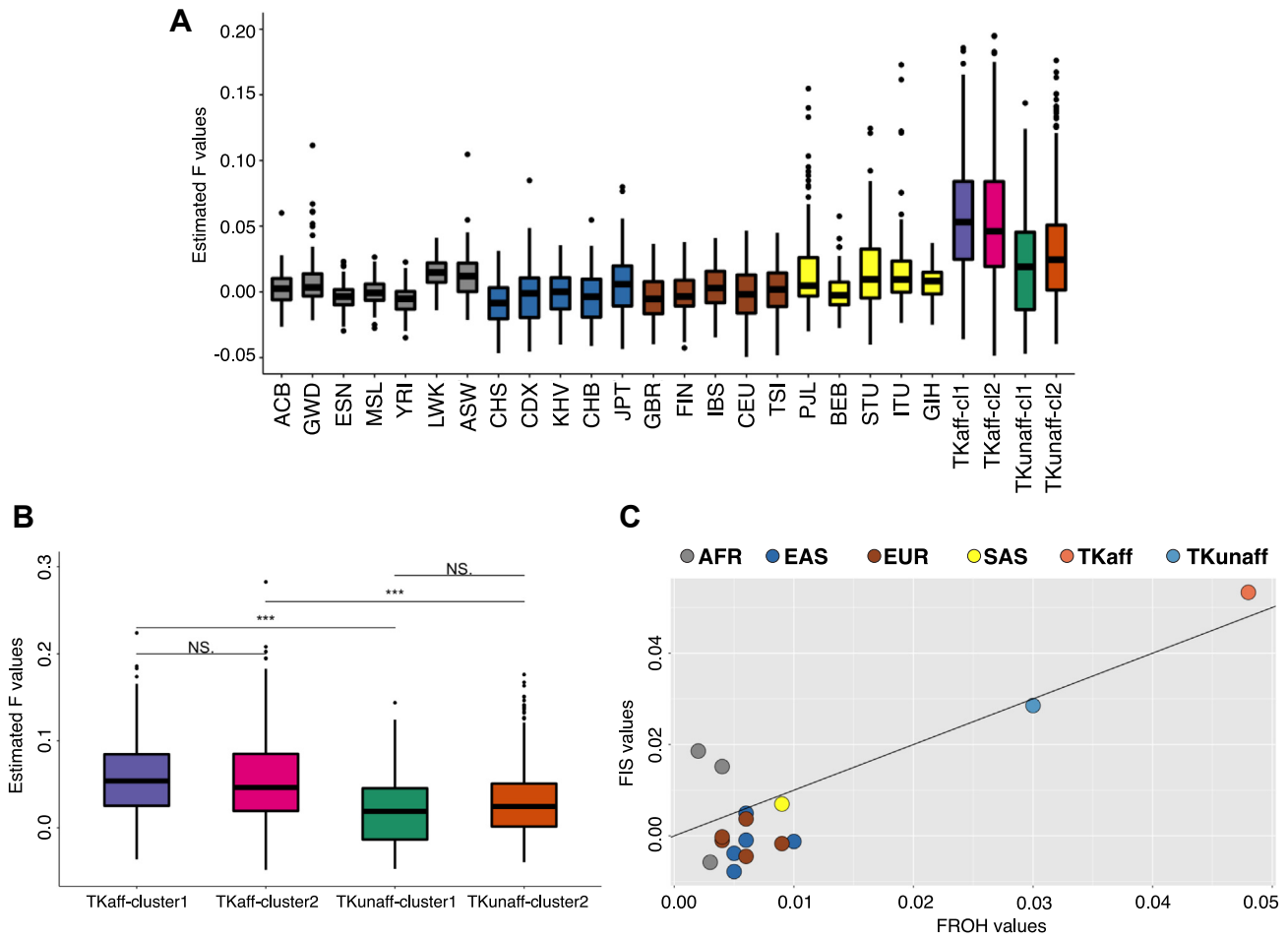

Figure 3 Higher estimated inbreeding coefficient F values in the TK individuals. A. The box plots report the estimated inbreeding coefficient levels (F) calculated from ES data. Both of the TK affected cluster 1 (purple), affected cluster 2 (dark pink), unaffected cluster 1 (dark green), and unaffected cluster 2 (dark orange) individuals showed higher F values on average compared with the African (AFR) (gray), East Asian (EAS) (dark blue), European (EUR) (brown), and South Asian (SAS) (yellow) subpopulations from the 1000 Genomes project. B. The box plots show the estimated F values of TK affected cluster 1 (TKaff-cluster1) and cluster 2 (TKaff-cluster2) individuals (mean = 0.055 vs 0.052) were significantly higher compared with the TK unaffected cluster 1 (TKunaff-cluster1) and cluster 2 (TKunaff-cluster2) individuals, respectively (Wilcoxon test one-tailed *P* values 2.3e-6 and 2.2e-16). There was no significant difference noted between F values of the TK affected cluster 1 and cluster 2 individuals (*P* value = 0.34) and the TK unaffected cluster 1 and cluster 2 (*P* = .11). *P* values indicated above each pair of groups compared (***$P < .001$). Outliers are not shown in the box plots. C. The scatter plot compares the average fraction of individual genome covered by long-sized (>2.64 Mb) ROHs (FROH) with the average estimated F values (FIS) for the African (AFR), East Asian (EAS), European (EUR), South Asian (SAS), and TK affected (TK-aff) and TK unaffected (TK-unaff) samples.

number of long-sized ROH regions that arose on recently configured "young" haplotypes because of recent parental relatedness. To test this, we identified first ROH regions from ES data using an informatics tool, BafCalculator (https://github.com/BCM-Lupskilab/BafCalculator)[32] that calculates genomic intervals with absence of heterozygosity from unphased ES data as a surrogate measure of ROH. To obtain copy-number-neutral ROH regions from the calculated ROH intervals, we excluded a subset of the apparent homozygous regions caused by common variant CNV deletions. Applying Gaussian-mixture modeling with the mclust R package, we classified the genomic intervals for ROH regions detected through BafCalculator into 3 length classes: short-sized (0.210-0.671 Mb), medium-sized (0.671-2.64 Mb), and long-sized ROHs (>2.64 Mb). To show that those ROH regions in these 3 different and discrete categories are not overlapping and are formed by different factors (eg, recent inbreeding and local recombination rate), we examined the distribution of those ROH regions of long-, medium-, and short-sized ROH regions and observed that they were located in different parts of the genome, ie, they map to different genetic locus intervals, and those regions were distributed non-uniformly across the genome as clearly visualized in the circos plot (Supplemental Figure 5). The non-uniform distribution of ROHs may be formed across the genome because of several physical properties of the human genome: they may involve some genes that are targets of

positive selection in a population,[57] or they may include small structural variant (SV) inversions that suppress recombination.[58,59]

We also examined the effect of local recombination rate on the formation of those ROHs of 3 different and discrete ROH size classes. In alignment with the finding that ROHs of long-sized, medium-sized, and short-sized ROHs distributed in different parts of the genome (Supplemental Figure 5), we found that long-sized ROHs are more likely to contain human genome recombination cold spot regions[60] compared with the medium-sized (Wilcoxon test one-tailed P value = 1.54e-13) and short-sized ROHs (Wilcoxon test one-tailed P value = 2.08e-29). Overall, our computational analyses showed that we uncovered 3 different and discrete ROH size classes that do not overlap with each other (distributed in different parts of the genome) and are specific and meaningful for the TK cohort. In addition, our data support the notion that "genome geography of recombination rates" (ie, positions of "coldspots" vs "hotspots" for recombination] may influence genetic architecture in specific populations.

To investigate the effect of inbreeding level on the genetic architecture of disease traits and to uncover which genomic regions are more contributory to disease phenotypes in the TK population, we compared the ROH length distribution between the TK affected and unaffected participants. As expected, the TK affected participants (with an average of estimated F values = 0.053) have a higher level of estimated F values compared with TK unaffected participants (with an average of estimated F values = 0.028), given that consanguinity is well established to be a risk factor for rare disorders. This increase in the estimated F levels in TK affected vs unaffected participants was reflected in an increase in the long-sized ROH total size (median = 111.71 Mb vs 34.41 Mb, Wilcoxon test one-tailed P value = 2.09e-18) and number (median = 13 vs 6, Wilcoxon test one-tailed P value = 6.14e-16) but not medium-sized and short-sized ROHs (Figure 4A, Supplemental Figure 6).

In parallel, we performed a correlation analysis to explore the hypothesis that higher estimated F values observed in the TK cohort manifested an increased genome-wide burden of long-sized (>2.64 Mb) ROHs. Our analysis revealed that as the estimated inbreeding levels increase in the TK affected participants, the genome-wide burden of long-sized ROHs increase (ρ = 0.83, Supplemental Figure 7A) but not the genome-wide burden of medium-sized ROHs (ρ = 0.29, Supplemental Figure 7B) or short-sized ROH segments (ρ = −0.25, Supplemental Figure 7C). This analysis supports the notion that recent inbreeding is most likely to contribute to the formation of long-sized ROHs but not to the other defined categories (ie, medium and short). Taken together, both analyses showed that, as the estimated inbreeding levels increase within the TK cohort (TK affected vs TK unaffected participants and within the TK affected participants), the genome-wide burden of long-sized ROH regions also increases.

We next examined specifically the characteristics of other ROH size categories. The total length of short-sized ROH regions, which result from short homozygous blocks on ancient haplotypes, showed a moderate negative correlation with the estimated F values (Supplemental Figure 7C). Concordant with the lower estimated F values in the TK unaffected participants, the short-sized ROH regions in the unaffected participants were greater in total size (median = 31 vs 28.62 Mb, Wilcoxon test one-tailed P value = 2.41e-7) and higher in number (median = 126 vs 122, Wilcoxon test one-tailed P value = 7.65e-8) than those of affected participants (Supplemental Figure 6B). In contrast, medium-sized ROH regions that arise mostly because of background relatedness[27,57] were present in a slightly higher number in the TK unaffected compared with those of affected participants (median = 34 vs 32, Wilcoxon test one-tailed P value = .04) but did not differ significantly in total length (median = 39.11 vs 39.77 Mb, Wilcoxon test one-tailed P value = .199) between the TK affected and unaffected participants (Supplemental Figure 6A).

Taken together, these ROH analyses support the notion that long-, medium-, and short-sized ROH regions result from individual or personal genome population history and dynamics, such as consanguinity level or background relatedness,[27,57] in nuclear families and clans. Of note, the TK affected participants show a significantly increased overall burden of ROH segments per genome compared with unaffected participants, stemming from an increased genome-wide burden of long-sized ROHs.

Enrichment of predicted deleterious variants in long-sized ROH regions

Then, we tested the hypothesis that long-sized ROH regions that arose because of recent parental relatedness would be enriched for rare homozygous deleterious variants because insufficient generational time would not yet have allowed for selective elimination of such detrimental alleles from a population.[61-63] We first identified rare homozygous deleterious variants (MAF $\leq$ 0.05) predicted to be potentially deleterious, by selecting variants above a CADD PHRED-scaled score of $\geq$15 and with a prediction tool algorithm, NMDescPredictor, to predict potential loss-of-function variants.[64] To account for variable variant rates across different samples and genomic regions, we computed a variant density metric by normalizing the count of rare homozygous deleterious variants to the count of rare homozygous synonymous variants. Because ROH regions are likely to enable deleterious variation to exist in a homozygous form, we then investigated the contribution of ROH regions to deleterious variant density in the TK cohort, by grouping rare homozygous deleterious variants into 4 groups based on their genomic location. As expected, in both the TK affected and unaffected participants, we observed an increased level of variant burden in ROH regions that was most striking in long-sized ROH regions,

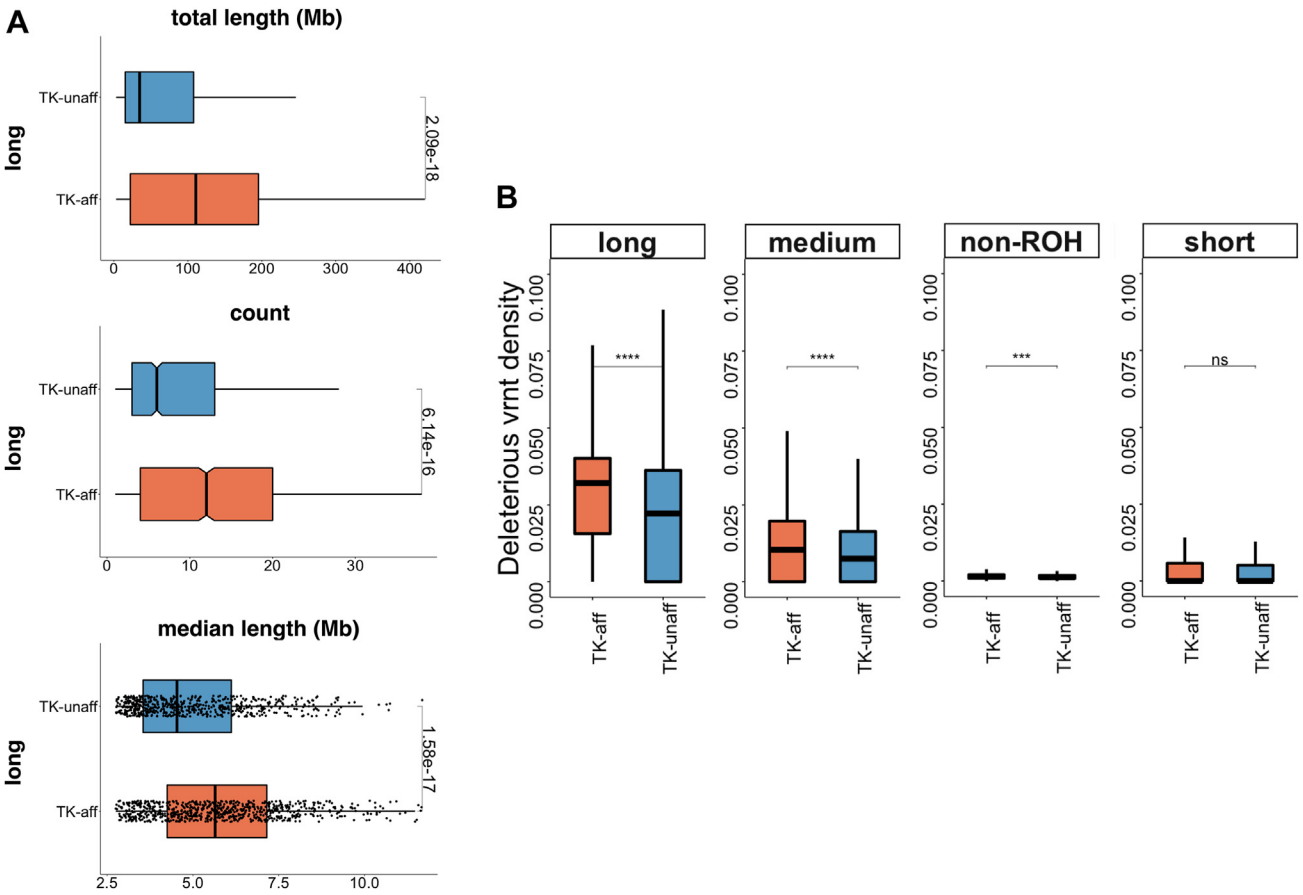

Figure 4 Long-sized ROH regions with increased density of rare deleterious variants are enriched in TK affected participants. A. The features of long-sized ROHs were displayed irrespective of their total size (Mb) (top panel), the number of ROH blocks (middle panel) and the median length of ROH blocks (Mb) (bottom panel). In each panel, horizontal box plots compare the 2 individual groups as the TK affected (TK-aff) and unaffected (TK-unaff) individuals. A one-sided Wilcoxon rank-sum test was used to test the difference in the TK-aff vs TK-unaff individuals and is indicated to the right of each panel. B. We calculated a variant density metric by normalizing the count of rare homozygous deleterious variants to the count of rare homozygous synonymous variants. The density of rare homozygous deleterious variants manifested in long-sized, medium-sized, short-sized ROH regions, and not ROH (non-ROH) regions in the TK affected (TK-aff) and TK unaffected (TK-unaff) participant subgroups. We then compared the rare homozygous deleterious variant density between the TK-aff and TK-unaff subgroups in each ROH size group and non-ROHs. *P* value significance levels were marked on the top of each pair of groups compared (***$P < .001$, ****$P < .0001$). Outliers were not shown in the box plots.

which are more likely to be shaped by young haplotype blocks including deleterious variation. This was followed by medium-sized ROH, short-sized ROH, and non-ROH regions (Figure 4B).

By comparison, we observed a significantly increased level of rare homozygous deleterious variant density in the TK affected compared with unaffected participants that was most strikingly outlined in long-sized (Wilcoxon test one-tailed *P* value = 3.35e-11) and to a lesser extent in medium-sized ROHs (Wilcoxon test one-tailed *P* value = 1.9e-9) and non-ROHs (Wilcoxon test one-tailed *P* value = 4.36e-2, Figure 4B). Short-sized ROHs do not seem to contribute to the overall rare homozygous deleterious variant density difference observed in ROH regions between the TK affected and unaffected participants (Wilcoxon test one-tailed *P* value = 1.75e-1, Figure 4B).

In summary, this genome analysis documented that long-sized ROHs were most enriched for rare homozygous deleterious variations compared with medium-sized, short-sized, and non-ROHs, irrespective of affection status in the TK cohort, indicating a contribution of recent parental relatedness to variant burden in the TK population (Figure 4B). This further demonstrated that the overall burden of rare homozygous deleterious variation was significantly increased in TK affected compared with unaffected participants and most prominent in long-sized ROHs, conceptually reflecting copy-number-neutral SV haplotypes derived by new variants and presenting with low rates of recombination in recent ancestors of the clan.

Phenotypic and population structure characterization of a mixed-disease cohort

The TK participants were recruited into the BHCMG because of suspected Mendelian disorders and presented with a wide

variety of clinical phenotypes. To define the range and spectrum of clinical phenotypes and the genetic traits observed in these individuals, we generated a phenotypic similarity score with the R package ontology. Similarity[29,47,48] and the Resnik approach to information content[65] to evaluate semantic similarity in a taxonomy between each pair of individuals using the HPO terms[66,67] recorded in PhenoDB. We performed an unsupervised clustering of those participants based on this phenotypic similarity score, revealing 5 major "disease phenotype groups" in the TK affected individuals. Clusters 1, 3, and 4 consist mainly of individuals with diverse phenotypic features, which do not fit into a "singular general disease category or group," ie, higher order HPO term or generalizable disease state or clinical diagnosis. Clusters 2 and 5 are formed of individuals who could be grouped by defined clinical entities largely reflective of complex disease traits, including the hypergonadotropic hypogonadism cohort[68] and the neurological disorders cohort,[4] respectively (Figure 5A).

Genotype-phenotype analyses revealing ROH-associated genetic architecture of diseases

We then tested the extent to which variant burden caused by rare homozygous deleterious variants in long-sized ROH regions explains clinical phenotypic features of the individuals in the TK cohort. To this end, we performed an unbiased genotype-phenotype correlation analysis based on the HPO terms.[69,70] In this analysis, we linked the disease genes present with rare coding homozygous and deleterious variation in an individual genome to their related HPO terms for the trait, as defined in the OMIM (https://www.omim.org) clinical synopsis, according to HPO annotation resource databases. Then, for the clinical phenotypic features of each patient recorded in PhenoDB,[54] we compared the associated HPO term sets to the merged HPO term sets for the disease genes defined for each ROH category, using a semantic similarity score metric that controls for the number of genes and ROH block size from a permutation approach.[29,47,48] These analyses showed that, in TK affected individuals, genes with rare homozygous deleterious variants in long-sized ROH regions are most informative for clinical phenotypic features objectively assessed and compared, from the information submitted to and collated within PhenoDB[54] compared with medium-sized ROHs (P value = 1.7e-1), short-sized ROHs (P value = 1.2e-3) and non-ROHs (P value = 5e-2) (Figure 5B).

These analyses also revealed that the top-ranking genes with rare coding homozygous deleterious variant alleles were located in the long-sized ROH regions in 152 patients (phenotypic similarity score ≥ 1). Importantly, 75 out of those 152 patients were found to carry rare coding homozygous and deleterious variation located in ≥2 OMIM disease genes that contribute significantly to the patients' phenotypes on the basis of phenotypic similarity score (score ≥ 1). In line with these findings, we indicate a subset of the TK affected participants in the cohort present with neurodevelopmental disorders (N =

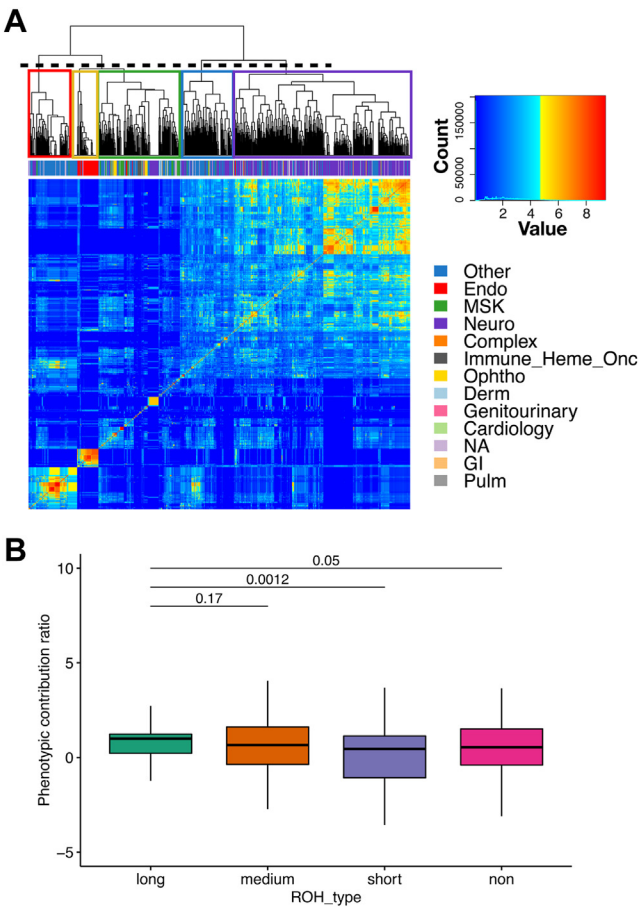

Figure 5 Contribution of variants in ROH to disease trait phenotype. A. A heatmap depicting the unsupervised hierarchical clustering of the TK affected patients into 5 general phenotypic categories or groups by calculating a pairwise phenotypic similarity score (clusters 1 [colored in red], 2 [colored in yellow], 3 [colored in green], 4 [colored in blue], and 5 [colored in purple]). B. The comparison between the long-sized ROH group with medium-sized, short-sized ROHs, and non-ROHs with regard to the performance of their variants in contribution to the disease trait phenotype presented in the TK patients. The y-axis describes the ratios of z-scores calculated for variants located in a specific ROH size group, eg, long-sized ROH group, vs z-scores calculated for all variants of interest. Z-scores were calculated for each tested disease gene list performing 1000 simulations. In each simulation, a permuted disease gene list was selected that has a same number of genes as the tested disease gene list. Then, a similarity score was computed between the patient's clinical features and the associated HPO terms of that permuted disease gene list.

234) that were analyzed by expert clinicians. In 176 of those 234 studied participants (75.2%), a plausible and genetically parsimonious molecular etiology due to rare coding variation was identified. Importantly, out of those 176 participants, 51 families (51/176 = 28.9%) were identified with multi-locus pathogenic variation, mostly driven by ROHs.[7] The retrospective analysis of those 176 participants revealed that cases with diagnoses involving ≥2 disease gene loci (N = 51) were found to have significantly higher F values (Wilcoxon test one-tailed P value = 9.35e-5, Figure 6A) and significantly

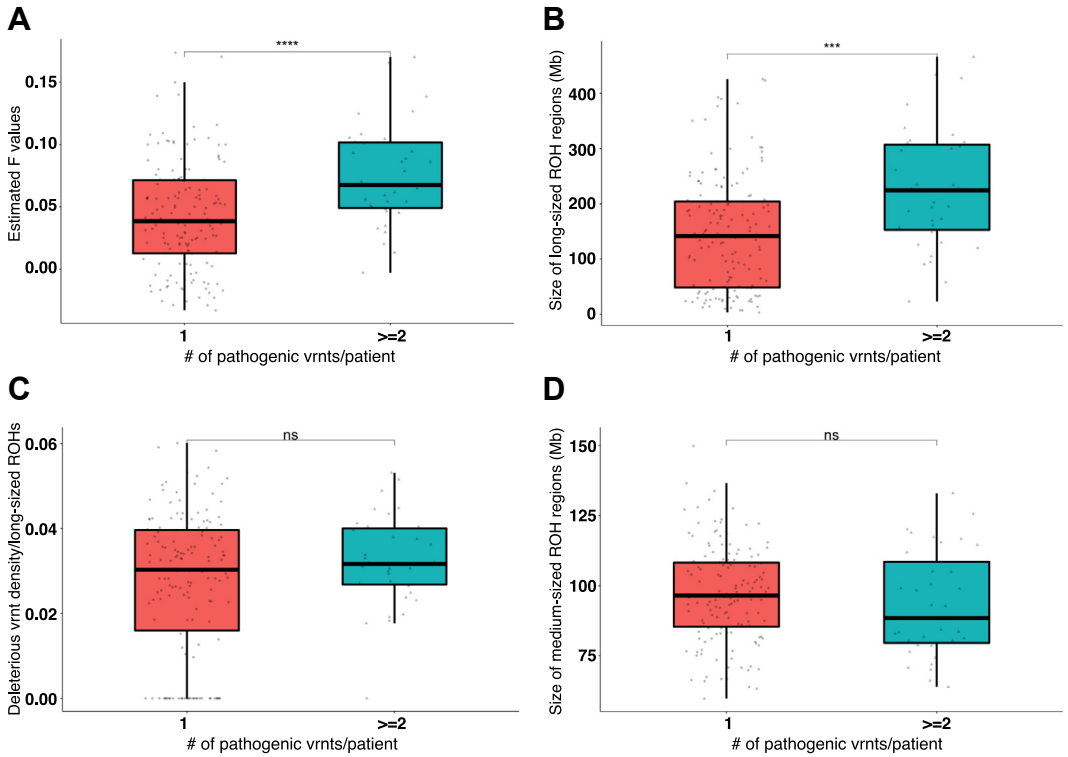

Figure 6 **An elevated variant burden in long-sized ROHs can produce blended phenotypes.** A retrospective analysis of previously published cases ($N = 217$) compared the cases with diagnoses involving ≥2 disease gene loci ($N = 40$) compared with cases with diagnoses involving 1 disease gene locus in terms of (A) estimated F values, (B) total span of long-sized ROHs, (C) deleterious variant density manifested in long-sized ROHs, and (D) total span of medium-sized ROHs. *P* values indicated above each pair of groups compared (***$P <$ .001, ****$P <$ .0001).

increased total span of long-sized ROHs (Wilcoxon test one-tailed *P* value = 1.06e-4, Figure 6B) compared with cases with diagnoses involving 1 disease gene locus. On the other hand, we found that there were no significant differences between these 2 groups of patients in terms of rare homozygous deleterious variant density in long-sized ROHs and total span of medium-sized ROHs (Wilcoxon test one-tailed *P* value = .223, Figure 6C and D). In summary, our results provide compelling evidence that an increased level of consanguinity observed in the TK affected compared with unaffected participants is correlated with a greater total span of long-sized ROH blocks and significantly increased level of rare homozygous deleterious variant density and thereby an elevated variant burden that likely contributes to the disease trait(s) phenotype observed in the individuals studied.

Discussion

Our results provide compelling evidence that an increased level of consanguinity observed in the TK affected compared with unaffected participants is correlated with a greater total span of long-sized ROH blocks and significantly increased level of rare homozygous deleterious variant density and thereby an elevated variant burden that likely contributes to rare disease trait(s). Comprehensive catalogs of common and rare human genetic variation stored in population variant databases, such as ExAC,[71] gnomAD[55] (https://gnomad.broadinstitute.org), and ARIC,[64,72] have proven to be powerful resources and interpretive tools to identify rare, ultra-rare, and potentially pathogenic variations that influence gene action and expression of Mendelian disease traits. However, many world populations remain for which genetic variation is underrepresented or entirely absent in such databases, including populations with more prevalent consanguinity and/or a considerable amount of admixture, such as the TK population.[41]

Beyond the elucidation of the biological basis and molecular pathogenesis of disease, the mapping of a locus at which variation might contribute to disease analysis and characterization of molecular features in personal genomes can potentially provide some insights into the genetic architecture contributing to disease traits in a population and evolution of personal genomes. From our TK population cohort, we show that investigation of the molecular features of ROH regions culled from personal unphased ES data through BafCalculator (https://github.com/BCM-Lupskilab/BafCalculator)[32] and classified through Gaussian-mixture machine learning modeling into 3-length ROH classes specific to the TK population can provide insights into the haplotype derivation and genetic architecture of disease.

Long-sized ROH (>2.64 Mb) regions that arose on recently configured haplotype blocks were greater in size (both total and median length) and number in the genomes of the TK affected compared with TK unaffected participants (median = 111.71 Mb vs 34.41 Mb, Wilcoxon test one-tailed *P* value = 2.09e-18 and median = 13 vs 6, Wilcoxon test one-tailed *P* value = 6.14e-16) without any significant cluster-specific differences (Figure 4A, Supplemental Figures 8 and 9). This finding is concordant with higher estimated F values (0.053 vs 0.028, Wilcoxon test one-tailed *P* value = 2.35e-18) and high level of consanguinity due to recent shared ancestors in the TK population.[56] Characterization of ROH genomic intervals also revealed that long-sized ROH regions on newly derived haplotypes are enriched particularly with rare homozygous deleterious variants specifically in the TK affected compared with TK unaffected participants (Wilcoxon test one-tailed *P* value = 3.35e-11). These findings support the notion that the inbreeding level results in an increase in the genome-wide burden of long-sized ROH along with an increase in their rare homozygous deleterious variation density (Figure 4A and B, Supplemental Figure 10).

To test the contribution of those ROH-derived rare variant combinations to Mendelian rare disease traits, we performed a large-scale analysis of genotype-phenotype correlations in the TK cohort from their clinical features captured as structured HPO terms[66,67] (https://hpo.jax.org/app/data/annotations) recorded in PhenoDB.[54] A systematic, integrative, and quantitative analysis revealed that the combinatorial phenotypic effect of ultra-rare variants embedded within long-size ROH regions strongly explain the observed rare disease trait(s) in the TK cohort. Corroborating these findings, a retrospective analysis of previously published cases ($N = 176$)[3,4,7,73] demonstrated that the cases with multi-locus pathogenic variation; $N = 51$) (diagnoses involving ≥2 disease gene loci) are more likely to have an increased genome-wide burden of long-sized ROH regions (Wilcoxon test one-tailed *P* value = 1.06e-4, Figure 6B), corresponding to higher estimated F values (Wilcoxon test one-tailed *P* value = 9.35e-5, Figure 6A) compared with cases with diagnoses involving 1 disease gene locus. That these variants were ultra-rare within the TK population itself suggests that they may represent new variant events within individual families or the clan, the most basic unit of a population. Population substructure ultimately driving formation of these newly configured—and unique—disease haplotypes can be rapidly brought to homozygosity through IBD: each shaped by recombination, characterized by new variant, and even more rare than their constituent alleles.[74]

This phenomenon of new variants as ultra-rare variants on a newly derived haplotype may provide important insights into the molecular etiology of disease traits and further amplify several prior observations including (1) the role of multilocus variation underlying some cases of apparent phenotypic expansion,[3,29,32,47] (2) homozygosity of ultra-rare pathogenic variation in *PRUNE* (HGNC:13420) in sect-driven population isolates[4,75,76] in stark contrast with the contribution of rare (but not ultra-rare) *CLP1* (HGNC:16999) founder alleles to microcephaly and neurodevelopmental disease,[77,78] (3) the seemingly now-common identification of genes for which biallelic variation can lead to disease traits previously categorized as strictly "dominant Mendelian loci"-traits,[7,79-81] and (4) the emergence of a founder allele arisen by identity by state in Steel syndrome of the geographically isolated population of the Commonwealth of Puerto Rico vs clan genomics derived *COL27A1* (HGNC:22986) pathogenic variation in the TK population.[37]

Our findings also highlight the advantage of merging per-locus variation with genomics (inherited vs de novo variants) for gleaning insights into the genetic architecture contributing to disease in a population. Multiple genetic changes could be brought together per locus to generate unique haplotype blocks. One example of this phenomenon is the observation of UPD manifest by a long tract of homozygosity on the entirety of chromosome 7, explaining both the short stature phenotype in addition to the expected clinical features of cystic fibrosis (CF, OMIM #219700) in a child.[51] Viewed from such a perspective, disease phenotypes resulting from UPD and epigenetic/imprinting diseases may benefit from haplotype phased genomes and long read technologies that differentiate methylated W-C bases.[82,83] Another example of per-locus genetic studies is the modulation of disease risk through gene expression and dosage effects of regulatory common variant (expression quantitative trait loci, eQTLs) haplotype configurations of coding pathogenic variants and CNVs.[5,84-88] The high variant rate of recurrent genomic deletions (eg, driven by a special type of variant mechanism, nonallelic homologous recombination [NAHR]) may make these loci a significant contributor to autosomal recessive disease trait loci in populations because of a compound heterozygous CNV + SNV allelic combination.[89] To apply a "merging" of genetics (per locus variation) and genomics (inherited "variome" and de novo variants) thinking to the clinic, our data emphasize identifying and systematically analyzing the ROH genomic intervals culled from the patient's personal unphased ES data with BafCalculator (https://github.com/BCM-Lupskilab/BafCalculator).[32] If the degree of parental relatedness as judged by the ROH size culled from unphased ES data suggests that the patient may come from a clan with a high estimated coefficient of consanguinity, rare homozygous variants mapping within the ROH regions and their combinatorial effect could be prioritized in molecular analyses.[3,29,32,47] Culling ROH regions from unphased clinical ES data with BafCalculator (https://github.com/ BCM-Lupskilab/BafCalculator)[32] may also enhance the discovery of pathogenic homozygous or hemizygous exonic CNV deletions arising on newly derived SV haplotypes in a clan homozygosed by IBD.[90-93]

In summary, the findings in this study support the Clan Genomics hypothesis,[94-96] which suggests that newly configured haplotypes resulting from recent variants play a role in Mendelian diseases. The rapid generation of disease haplotypes driven by population substructure and shaped by recombination contributes to the occurrence of these rare

disease-causing haplotypes. That these haplotypes are even more rare than their constituent alleles reinforces the importance of newly arisen variants in rare recessive disease traits.

Data Availability

All variants reported herein have been aggregated within a TK variome database that is publicly available as a research and molecular diagnostic community resource (https://turkishvariomedb.shinyapps.io/tvdb/). The code generated during this study is available at (https://github.com/BCM-Lupskilab/BafCalculator). Exome variant data have been deposited to dbGaP (study accession phs000711.v5.p1) and/or AnVIL for all cases for which written, informed consent for sharing of data through controlled-access databases has been obtained, in keeping with our IRB and BCM protocol H-29697. Requests for further information on raw data, genomic and phenotypic analyses, and DNA samples may be directed to, and will be fulfilled by the lead contact James R. Lupski (jlupski@bcm.edu).

Acknowledgments

The authors thank all patients, their families, and their referring physicians who submitted samples for genomic studies. No additional compensation was offered for these studies.

Funding

This work was supported in part by the US National Human Genome Research Institute (NHGRI)/National Heart Lung and Blood Institute (NHLBI) grant number UM1HG006542 to the Baylor Hopkins Center for Mendelian Genomics (BHCMG), the US National Human Genome Research Institute (NHGRI) U01HG011758 to the Baylor College of Medicine for the Genomics Research to Elucidate the Genetics of Rare Disease consortium (BCM-GREGoR), the National Institute of Neurological Disorders and Stroke (NINDS) R35NS105078, and the National Human Genome Research Institute U54-HG003273. J.E.P. was supported by NHGRI K08 HG008986.

Author Information

Conceptualization: Z.C.-A., X.S., F.C.C., D.P., J.R.L.; Data Curation: Z.C.-A., X.S., F.C.C., T.G., T.B.-Y.; Formal Analysis: Z.C.-A., X.S., F.C.C., D.P., J.R.L.; Supervision: J.R.L., D.V.; Project Administration: S.N.J., T.G., C.M.B.C., D.M.M., P.L., E.B., R.A.G., C.A.S., J.E.P.; Funding Acquisition: R.A.G., J.E.P., D.V., J.R.L.; Clinical Resources: D.P., E.K., Y.B., T.M., A.H., V.R.S., R.A.L., N.S.; Writing-original draft: Z.C.-A., X.S., F.C.C., D.P., J.R.L.; Writing-review and editing: Z.C.-A., X.S., F.C.C., D.P., R.A.L., C.A.S., C.M.B.C., J.E.P., J.R.L. All contributing co-authors have read, edited, and approved the final manuscript.

Ethics Declaration

This study was approved by the Institutional Review Board at Baylor College of Medicine (IRB protocol # H-29697). The authors recruited 1416 unrelated TK individuals (669 females and 747 males) in the Baylor Hopkins Center for Mendelian Genomics (BHCMG) cohort (data freeze: December 2011-October 2020) after all relevant subjects or legally authorized representatives provided written informed consent for the use of their DNA and personal genomes for the identification of potential disease-contributing variants and for broad data sharing.

Conflict of Interest

James R. Lupski has stock ownership in 23andMe and is a paid consultant for Genomics International. The Department of Molecular and Human Genetics at Baylor College of Medicine derives revenue from molecular genetic and genomic testing offered at BG (http://www.bcm.edu/geneticlabs/). James R. Lupski serves on the SAB of BG. All other authors declare no conflicts of interest.

Additional Information

The online version of this article (https://doi.org/10.1016/j.gimo.2024.101830) contains supplemental material, which is available to authorized users.

Web Resources

1000 Genomes Project Database, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
Atherosclerosis Risk in Communities Study (ARIC) Database, http://www2.cscc.unc.edu/aric/
BafCalculator ROH Detection Tool, https://github.com/BCM-Lupskilab/BafCalculator
dbGaP, https://www.ncbi.nlm.nih.gov/gap
ExAC Browser, http://exac.broadinstitute.org/
gnomAD Browser, https://gnomad.broadinstitute.org/
Human Genome Diversity Panel, https://www.internationalgenome.org/data-portal/data-collection/hgdp
Human Phenotype Ontology (HPO) Terms Annotation, https://hpo.jax.org/app/data/annotations
Mercury Analysis Pipeline, https://www.hgsc.bcm.edu/software/mercury

NMDescPredictor, https://nmdprediction.shinyapps.io/nmdescpredictor/
NMD Escape Intolerance Score, https://nmdprediction.shinyapps.io/nmdescintolerancescore/
OMIM, http://www.omim.org/
PhenoDB, https://phenodb.org/
The Greater Middle Eastern Variome Database, https://annovar.openbioinformatics.org/en/latest/user-guide/filter/
Turkish Variome Database, https://turkishvariomedb.shinyapps.io/tvdb/

Affiliations

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; [2]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX; [3]Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL; [4]Instituto de Salud Carlos III, National Center of Microbiology, Madrid, Spain; [5]Section of Neurology, Department of Pediatrics, Baylor College of Medicine, Houston, TX; [6]Department of Pathology, Baylor University Medical Center, Dallas, TX; [7]Texas A&M School of Medicine, Texas A&M University, Dallas, TX; [8]Division of Genomic Diagnostics, Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA; [9]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [10]Institute of Computer Science, Warsaw University of Technology, Warsaw, Poland; [11]Department of Biostatistics and Bioinformatics, Institute of Health Sciences, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey; [12]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; [13]Department of Pediatrics, Baylor College of Medicine, Houston, TX; [14]Department of Ophthalmology, Cullen Eye Institute, Baylor College of Medicine, Houston, TX; [15]McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; [16]Texas Children's Hospital, Houston, TX; [17]Pacific Northwest Research Institute, Seattle, WA; [18]Baylor Genetics, Houston, TX

References

1. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet*. 2001;17(9):502-510. http://doi.org/10.1016/s0168-9525(01)02410-6
2. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*. 2005;6(2):109-118. http://doi.org/10.1038/nrg1522
3. Pehlivan D, Bayram Y, Gunes N, et al. The genomics of arthrogryposis, a complex trait: candidate genes and further evidence for oligogenic inheritance. *Am J Hum Genet*. 2019;105(1):132-150. http://doi.org/10.1016/j.ajhg.2019.05.015
4. Karaca E, Harel T, Pehlivan D, et al. Genes that affect brain structure and function identified by rare variant analyses of Mendelian neurologic disease. *Neuron*. 2015;88(3):499-513. http://doi.org/10.1016/j.neuron.2015.09.048
5. Wu N, Ming X, Xiao J, et al. *TBX6* null variants and a common hypomorphic allele in congenital scoliosis. *N Engl J Med*. 2015;372(4):341-350. http://doi.org/10.1056/NEJMoa1406829
6. Groopman EE, Povysil G, Goldstein DB, Gharavi AG. Rare genetic causes of complex kidney and urological diseases. *Nat Rev Nephrol*. 2020;16(11):641-656. http://doi.org/10.1038/s41581-020-0325-2
7. Mitani T, Isikay S, Gezdirici A, et al. High prevalence of multilocus pathogenic variation in neurodevelopmental disorders in the Turkish population. *Am J Hum Genet*. 2021;108(10):1981-2005. http://doi.org/10.1016/j.ajhg.2021.08.009
8. Alkuraya FS. Homozygosity mapping: one more tool in the clinical geneticist's toolbox. *Genet Med*. 2010;12(4):236-239. http://doi.org/10.1097/GIM.0b013e3181ceb95d
9. Alkuraya FS. Autozygome decoded. *Genet Med*. 2010;12(12):765-771. http://doi.org/10.1097/GIM.0b013e3181fbfcc4
10. Alkuraya FS. How the human genome transformed study of rare diseases. *Nature*. 2021;590(7845):218-219. http://doi.org/10.1038/d41586-021-00294-7
11. Torkamani A, Pham P, Libiger O, et al. Clinical implications of human population differences in genome-wide rates of functional genotypes. *Front Genet*. 2012;3:211. http://doi.org/10.3389/fgene.2012.00211
12. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337(6090):64-69. http://doi.org/10.1126/science.1219240
13. Lohmueller KE, Indap AR, Schmidt S, et al. Proportionally more deleterious genetic variation in European than in African populations. *Nature*. 2008;451(7181):994-997. http://doi.org/10.1038/nature06611
14. Clark DW, Okada Y, Moore KHS, et al. Associations of autozygosity with a broad range of human phenotypes. *Nat Commun*. 2019;10(1):4957. http://doi.org/10.1038/s41467-019-12283-6
15. Joshi PK, Esko T, Mattsson H, et al. Directional dominance on stature and cognition in diverse human populations. *Nature*. 2015;523(7561):459-462. http://doi.org/10.1038/nature14618
16. Keller MC, Simonson MA, Ripke S, et al. Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS Genet*. 2012;8(4):e1002656. http://doi.org/10.1371/journal.pgen.1002656
17. Yengo L, Wray NR, Visscher PM. Extreme inbreeding in a European ancestry sample from the contemporary UK population. *Nat Commun*. 2019;10(1):3719. http://doi.org/10.1038/s41467-019-11724-6
18. Smith CAB. The detection of linkage in human genetics. *J R Stat Soc Series B Stat Methodol*. 1953;15(2):153-184. http://doi.org/10.1111/j.2517-6161.1953.tb00133.x
19. Morton NE. Genetic epidemiology of hearing impairment. *Ann N Y Acad Sci*. 1991;630:16-31. http://doi.org/10.1111/j.1749-6632.1991.tb19572.x
20. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*. 1987;236(4808):1567-1570. http://doi.org/10.1126/science.2884728
21. Broman KW, Weber JL. Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. *Am J Hum Genet*. 1999;65(6):1493-1500. http://doi.org/10.1086/302661
22. Seelow D, Schuelke M, Hildebrandt F, Nürnberg P. HomozygosityMapper—an interactive approach to homozygosity mapping. *Nucleic Acids Res*. 2009;37(Web Server issue):W593-W599. http://doi.org/10.1093/nar/gkp369
23. Houwen RH, Baharloo S, Blankenship K, et al. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat Genet*. 1994;8(4):380-386. http://doi.org/10.1038/ng1294-380
24. Alazami AM, Patel N, Shamseldin HE, et al. Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Rep*. 2015;10(2):148-161. http://doi.org/10.1016/j.celrep.2014.12.015

25. Alazami AM, Shaheen R, Alzahrani F, et al. *FREM1* mutations cause bifid nose, renal agenesis, and anorectal malformations syndrome. *Am J Hum Genet*. 2009;85(3):414-418. http://doi.org/10.1016/j.ajhg.2009.08.010
26. Alazami AM, Al-Saif A, Al-Semari A, et al. Mutations in *C2orf37*, encoding a nucleolar protein, cause hypogonadism, alopecia, diabetes mellitus, mental retardation, and extrapyramidal syndrome. *Am J Hum Genet*. 2008;83(6):684-691. http://doi.org/10.1016/j.ajhg.2008.10.018
27. McQuillan R, Leutenegger AL, Abdel-Rahman R, et al. Runs of homozygosity in European populations. *Am J Hum Genet*. 2008;83(3):359-372. http://doi.org/10.1016/j.ajhg.2008.08.007
28. Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. Genomic runs of homozygosity record population history and consanguinity. *PLoS One*. 2010;5(11):e13996. http://doi.org/10.1371/journal.pone.0013996
29. Posey JE, Harel T, Liu P, et al. Resolution of disease phenotypes resulting from multilocus genomic variation. *N Engl J Med*. 2017;376(1):21-31. http://doi.org/10.1056/NEJMoa1516767
30. Kaiser VB, Svinti V, Prendergast JG, et al. Homozygous loss-of-function variants in European cosmopolitan and isolate populations. *Hum Mol Genet*. 2015;24(19):5464-5474. http://doi.org/10.1093/hmg/ddv272
31. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet*. 2018;19(4):220-234. http://doi.org/10.1038/nrg.2017.109
32. Karaca E, Posey JE, Coban Akdemir Z, et al. Phenotypic expansion illuminates multilocus pathogenic variation. *Genet Med*. 2018;20(12):1528-1537. http://doi.org/10.1038/gim.2018.33
33. Katsanis N, Ansley SJ, Badano JL, et al. Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. *Science*. 2001;293(5538):2256-2259. http://doi.org/10.1126/science.1063525
34. Bejjani BA, Lewis RA, Tomey KF, et al. Mutations in *CYP1B1*, the gene for cytochrome P4501B1, are the predominant cause of primary congenital glaucoma in Saudi Arabia. *Am J Hum Genet*. 1998;62(2):325-333. http://doi.org/10.1086/301725
35. Bejjani BA, Stockton DW, Lewis RA, et al. Multiple *CYP1B1* mutations and incomplete penetrance in an inbred population segregating primary congenital glaucoma suggest frequent de novo events and a dominant modifier locus. *Hum Mol Genet*. 2000;9(3):367-374. http://doi.org/10.1093/hmg/9.3.367
36. Gonzaga-Jauregui C, Harel T, Gambin T, et al. Exome sequence analysis suggests that genetic burden contributes to phenotypic variability and complex neuropathy. *Cell Rep*. 2015;12(7):1169-1183. http://doi.org/10.1016/j.celrep.2015.07.023
37. Gonzaga-Jauregui C, Yesil G, Nistala H, et al. Functional biology of the Steel syndrome founder allele and evidence for clan genomics derivation of *COL27A1* pathogenic alleles worldwide. *Eur J Hum Genet*. 2020;28(9):1243-1264. http://doi.org/10.1038/s41431-020-0632-x
38. Hashmi MA. Frequency of consanguinity and its effect on congenital malformation—a hospital based study. *J Pak Med Assoc*. 1997;47(3):75-78.
39. Bittles AH, Black ML. Evolution in health and medicine Sackler colloquium: consanguinity, human evolution, and complex diseases. *Proc Natl Acad Sci U S A*. 2010;107(suppl 1):1779-1786. http://doi.org/10.1073/pnas.0906079106
40. Bittles A. Consanguinity and its relevance to clinical genetics. *Clin Genet*. 2001;60(2):89-98. http://doi.org/10.1034/j.1399-0004.2001.600201.x
41. Tunçbílek E, Koc I. Consanguineous marriage in Turkey and its impact on fertility and mortality. *Ann Hum Genet*. 1994;58(4):321-329. http://doi.org/10.1111/j.1469-1809.1994.tb00729.x
42. Reid JG, Carroll A, Veeraraghavan N, et al. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics*. 2014;15:30. http://doi.org/10.1186/1471-2105-15-30
43. Challis D, Yu J, Evani US, et al. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics*. 2012;13:8. http://doi.org/10.1186/1471-2105-13-8
44. Farek J, Hughes D, Salerno W, et al. xAtlas: scalable small variant calling across heterogeneous next-generation sequencing experiments. *Gigascience*. 2022;12:giac125. http://doi.org/10.1093/gigascience/giac125
45. Bainbridge MN, Wiszniewski W, Murdock DR, et al. Whole-genome sequencing for optimized patient management. *Sci Transl Med*. 2011;3:87re83. http://doi.org/10.1126/scitranslmed.3002243
46. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. http://doi.org/10.1093/nar/gkq603
47. Liu P, Meng L, Normand EA, et al. Reanalysis of clinical exome sequencing Data. *N Engl J Med*. 2019;380(25):2478-2480. http://doi.org/10.1056/NEJMc1812033
48. James RA, Campbell IM, Chen ES, et al. A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. *Genome Med*. 2016;8(1):13. http://doi.org/10.1186/s13073-016-0261-8
49. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5(4):557-572. http://doi.org/10.1093/biostatistics/kxh008
50. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015;12(2):115-121. http://doi.org/10.1038/nmeth.3252
51. Spence JE, Perciaccante RG, Greig GM, et al. Uniparental disomy as a mechanism for human genetic disease. *Am J Hum Genet*. 1988;42(2):217-226.
52. Fromer M, Purcell SM. Using XHMM software to detect copy number variation in whole-exome sequencing data. *Curr Protoc Hum Genet*. 2014;81:7.23.1-7.23.21. http://doi.org/10.1002/0471142905.hg0723s81
53. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics*. 2014;47:11.12.1-11.1234. http://doi.org/10.1002/0471250953.bi1112s47
54. Sobreira N, Schiettecatte F, Boehm C, Valle D, Hamosh A. New tools for Mendelian disease gene identification: PhenoDB variant analysis module; and GeneMatcher, a web-based tool for linking investigators with an interest in the same gene. *Hum Mutat*. 2015;36(4):425-431. http://doi.org/10.1002/humu.22769
55. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. http://doi.org/10.1038/s41586-020-2308-7
56. Scott EM, Halees A, Itan Y, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat Genet*. 2016;48(9):1071-1076. http://doi.org/10.1038/ng.3592
57. Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet*. 2012;91(2):275-292. http://doi.org/10.1016/j.ajhg.2012.06.014
58. Curtis D, Vine AE, Knight J. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Ann Hum Genet*. 2008;72(2):261-278. http://doi.org/10.1111/j.1469-1809.2007.00411.x
59. Porubsky D, Höps W, Ashraf H, et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*. 2022;185(11):1986-2005.e26. http://doi.org/10.1016/j.cell.2022.04.017
60. Hussin JG, Hodgkinson A, Idaghdour Y, et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat Genet*. 2015;47(4):400-404. http://doi.org/10.1038/ng.3216
61. Pemberton TJ, Szpiech ZA. Relationship between deleterious variation, genomic autozygosity, and disease risk: insights from the 1000 genomes project. *Am J Hum Genet*. 2018;102(4):658-675. http://doi.org/10.1016/j.ajhg.2018.02.013
62. Szpiech ZA, Xu J, Pemberton TJ, et al. Long runs of homozygosity are enriched for deleterious variation. *Am J Hum Genet*. 2013;93(1):90-102. http://doi.org/10.1016/j.ajhg.2013.05.003

63. Szpiech ZA, Mak ACY, White MJ, et al. Ancestry-dependent enrichment of deleterious homozygotes in runs of homozygosity. *Am J Hum Genet*. 2019;105(4):747-762. http://doi.org/10.1016/j.ajhg.2019.08.011
64. Coban-Akdemir Z, White JJ, Song X, et al. Identifying genes whose mutant transcripts cause dominant disease traits by potential gain-of-function alleles. *Am J Hum Genet*. 2018;103(2):171-187. http://doi.org/10.1016/j.ajhg.2018.06.009
65. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. 1995:448-453.
66. Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*. 2014;42(Database issue):D966-D974. http://doi.org/10.1093/nar/gkt1026
67. Groza T, Köhler S, Moldenhauer D, et al. The human phenotype ontology: semantic unification of common and rare disease. *Am J Hum Genet*. 2015;97(1):111-124. http://doi.org/10.1016/j.ajhg.2015.05.020
68. Jolly A, Bayram Y, Turan S, et al. Exome sequencing of a primary ovarian insufficiency cohort reveals common molecular etiologies for a spectrum of disease. *J Clin Endocrinol Metab*. 2019;104(8):3049-3067. http://doi.org/10.1210/jc.2019-00248
69. Zhang C, Jolly A, Shayota BJ, et al. Novel pathogenic variants and quantitative phenotypic analyses of Robinow syndrome: WNT signaling perturbation and phenotypic variability. *HGG Adv*. 2022;3(1):100074. http://doi.org/10.1016/j.xhgg.2021.100074
70. Herman I, Jolly A, Du H, et al. Quantitative dissection of multilocus pathogenic variation in an Egyptian infant with severe neurodevelopmental disorder resulting from multiple molecular diagnoses. *Am J Med Genet A*. 2022;188(3):735-750. http://doi.org/10.1002/ajmg.a.62565
71. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291. http://doi.org/10.1038/nature19057
72. Gambin T, Jhangiani SN, Below JE, et al. Secondary findings and carrier test frequencies in a large multiethnic sample. *Genome Med*. 2015;7(1):54. http://doi.org/10.1186/s13073-015-0171-1
73. Bayram Y, Karaca E, Coban Akdemir Z, et al. Molecular etiology of arthrogryposis in multiple families of mostly Turkish origin. *J Clin Invest*. 2016;126(2):762-778. http://doi.org/10.1172/JCI84457
74. Narasimhan VM, Rahbari R, Scally A, et al. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat Commun*. 2017;8:303. http://doi.org/10.1038/s41467-017-00323-y
75. Zollo M, Ahmed M, Ferrucci V, et al. *PRUNE* is crucial for normal brain development and mutated in microcephaly with neurodevelopmental impairment. *Brain*. 2017;140(4):940-952. http://doi.org/10.1093/brain/awx014
76. Nistala H, Dronzek J, Gonzaga-Jauregui C, et al. NMIHBA results from hypomorphic *PRUNE1* variants that lack short-chain exopolyphosphatase activity. *Hum Mol Genet*. 2021;29(21):3516-3531. http://doi.org/10.1093/hmg/ddaa237
77. Karaca E, Weitzer S, Pehlivan D, et al. Human *CLP1* mutations alter tRNA biogenesis, affecting both peripheral and central nervous system function. *Cell*. 2014;157(3):636-650. http://doi.org/10.1016/j.cell.2014.02.058
78. Schaffer AE, Eggens VR, Caglayan AO, et al. *CLP1* founder mutation links tRNA splicing and maturation to cerebellar development and neurodegeneration. *Cell*. 2014;157(3):651-663. http://doi.org/10.1016/j.cell.2014.03.049
79. Yuan B, Pehlivan D, Karaca E, et al. Global transcriptional disturbances underlie Cornelia de Lange syndrome and related phenotypes. *J Clin Invest*. 2015;125(2):636-651. http://doi.org/10.1172/JCI77435
80. Rainger J, Pehlivan D, Johansson S, et al. Monoallelic and biallelic mutations in *MAB21L2* cause a spectrum of major eye malformations. *Am J Hum Genet*. 2014;94(6):915-923. http://doi.org/10.1016/j.ajhg.2014.05.005
81. Monies D, Abouelhoda M, Assoum M, et al. Lessons learned from large-scale, first-tier clinical exome sequencing in a highly consanguineous population. *Am J Hum Genet*. 2019;104(6):1182-1201. http://doi.org/10.1016/j.ajhg.2019.04.011
82. Carvalho CMB, Coban-Akdemir Z, Hijazi H, et al. Interchromosomal template-switching as a novel molecular mechanism for imprinting perturbations associated with Temple syndrome. *Genome Med*. 2019;11(1):25. http://doi.org/10.1186/s13073-019-0633-y
83. Lupski JR, Liu P, Stankiewicz P, Carvalho CMB, Posey JE. Clinical genomics and contextualizing genome variation in the diagnostic laboratory. *Expert Rev Mol Diagn*. 2020;20(10):995-1002. http://doi.org/10.1080/14737159.2020.1826312
84. Castel SE, Cervera A, Mohammadi P, et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet*. 2018;50(9):1327-1334. http://doi.org/10.1038/s41588-018-0192-y
85. Yang N, Wu N, Zhang L, et al. *TBX6* compound inheritance leads to congenital vertebral malformations in humans and mice. *Hum Mol Genet*. 2019;28(4):539-547. http://doi.org/10.1093/hmg/ddy358
86. Liu J, Wu N. Deciphering Disorders Involving Scoliosis and COmorbidities (DISCO) study, et al. *TBX6*-associated congenital scoliosis (TACS) as a clinically distinguishable subtype of congenital scoliosis: further evidence supporting the compound inheritance and *TBX6* gene dosage model. *Genet Med*. 2019;21(7):1548-1558. http://doi.org/10.1038/s41436-018-0377-x
87. Ren X, Yang N, Wu N, et al. Increased *TBX6* gene dosages induce congenital cervical vertebral malformations in humans and mice. *J Med Genet*. 2020;57(6):371-379. http://doi.org/10.1136/jmedgenet-2019-106333
88. Duan R, Hijazi H, Gulec EY, et al. Developmental genomics of limb malformations: allelic series in association with gene dosage effects contribute to the clinical variability. *HGG Adv*. 2022;3(4):100132. http://doi.org/10.1016/j.xhgg.2022.100132
89. Yuan B, Schulze KV, Assia Batzir N, et al. Sequencing individual genomes with recurrent genomic disorder deletions: an approach to characterize genes for autosomal recessive rare disease traits. *Genome Med*. 2022;14(1):113. http://doi.org/10.1186/s13073-022-01113-y
90. Gambin T, Akdemir ZC, Yuan B, et al. Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res*. 2017;45(4):1633-1648. http://doi.org/10.1093/nar/gkw1237
91. Yuan B, Wang L, Liu P, et al. CNVs cause autosomal recessive genetic diseases with or without involvement of SNV/indels. *Genet Med*. 2020;22(10):1633-1641. http://doi.org/10.1038/s41436-020-0864-8
92. Dharmadhikari AV, Ghosh R, Yuan B, et al. Copy number variant and runs of homozygosity detection by microarrays enabled more precise molecular diagnoses in 11,020 clinical exome cases. *Genome Med*. 2019;11(1):30. http://doi.org/10.1186/s13073-019-0639-5
93. Karolak JA, Vincent M, Deutsch G, et al. Complex compound inheritance of lethal lung developmental disorders due to disruption of the TBX-FGF pathway. *Am J Hum Genet*. 2019;104(2):213-228. http://doi.org/10.1016/j.ajhg.2018.12.010
94. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. Clan genomics and the complex architecture of human disease. *Cell*. 2011;147(1):32-43. http://doi.org/10.1016/j.cell.2011.09.008
95. Lupski JR. Clan genomics: from OMIM phenotypic traits to genes and biology. *Am J Med Genet A*. 2021;185(11):3294-3313. http://doi.org/10.1002/ajmg.a.62434
96. Lupski JR. Biology in balance: human diploid genome integrity, gene dosage, and genomic medicine. *Trends Genet*. 2022;38(6):554-571. http://doi.org/10.1016/j.tig.2022.03.001